Exhibit 20

neither ensures unbiased effect estimates nor follows from the requirement of unbiasedness. After demonstrating, by examples, the absence of logical connections between the statistical and the causal notions of confounding, we will define a stronger notion of unbiasedness, called "stable" unbiasedness, relative to which a modified statistical criterion will be shown necessary and sufficient. The necessary part will then yield a practical test for stable unbiasedness that, remarkably, does not require knowledge of all potential confounders in a problem. Finally, we will argue that the prevailing practice of substituting statistical criteria for the effect-based definition of confounding is not entirely misguided, because stable unbiasedness is in fact (i) what investigators have been (and perhaps should be) aiming to achieve and (ii) what statistical criteria can test.

6.2.2 Causal and Associational Definitions

In order to facilitate the discussion, we shall first cast the causal and statistical definitions of no-confounding in mathematical forms.[11]

Definition 6.2.1 (No-Confounding; Causal Definition)
Let M be a causal model of the data-generating process – that is, a formal description of how the value of each observed variable is determined. Denote by $P(y \mid do(x))$ the probability of the response event $Y = y$ under the hypothetical intervention $X = x$, calculated according to M. We say that X and Y are not confounded in M if and only if

$$P(y \mid do(x)) = P(y \mid x), \text{ or } P(x \mid do(y)) = P(x \mid y) \tag{6.10}$$

for all x and y in their respective domains, where $P(y \mid x)$ is the conditional probability generated by M. If (6.10) holds, we say that $P(y \mid x)$ is unbiased.

For the purpose of our discussion here, we take this causal definition as the meaning of the expression "no confounding." The probability $P(y \mid do(x))$ was defined in Chapter 3 (Definition 3.2.1, also abbreviated $P(y \mid \hat{x})$); it may be interpreted as the conditional probability $P^*(Y = y \mid X = x)$ corresponding to a controlled experiment in which X is randomized. We recall that this probability can be calculated from a causal model M either directly, by simulating the intervention $do(X = x)$, or (if $P(x, s) > 0$) via the adjustment formula (equation (3.19))

$$P(y \mid do(x)) = \sum_s P(y \mid x, s)\, P(s),$$

where S stands for any set of variables, observed as well as unobserved, that satisfy the back-door criterion (Definition 3.3.1). Equivalently, $P(y \mid do(x))$ can be written $P(Y(x) = y)$, where $Y(x)$ is the potential-outcome variable as defined in (3.51) or in

[11] For simplicity, we will limit our discussion to unadjusted confounding; extensions involving measurement of auxiliary variables are straightforward and can be obtained from Section 3.3. We also use the abbreviated expression "X and Y are not confounded," though "the effect of X on Y is not confounded" is more exact.

Rubin (1974). We bear in mind that the operator *do* $(\cdot)$, and hence also effect estimates and confounding, must be defined relative to a specific causal or data-generating model M because these notions are not statistical in character and cannot be defined in terms of joint distributions.

Definition 6.2.2 (No-Confounding; Associational Criterion)
Let T be the set of variables in a problem that are not affected by X. We say that X and Y are not confounded in the presence of T if each member Z of T satisfies at least one of the following conditions:

(U_1) *Z is not associated with X (i.e., $P(x \mid z) = P(x)$);*
(U_2) *Z is not associated with Y, conditional on X (i.e., $P(y \mid z, x) = P(y \mid x)$).*

Conversely, X and Y are said to be confounded if any member Z of T violates both (U_1) and (U_2).

Note that the associational criterion in Definition 6.2.2 is not purely statistical in that it invokes the predicate "affected by," which is not discernible from probabilities but rests instead on causal information. This exclusion of variables that are affected by treatments (or exposures) is unavoidable and has long been recognized as a necessary judgmental input to every analysis of treatment effect in observational and experimental studies alike (Cox 1958, p. 48; Greenland and Neutra 1980). We shall assume throughout that investigators possess the knowledge required for distinguishing variables that are affected by the treatment X from those that are not. We shall then explore what additional causal knowledge is needed, if any, for establishing a test of confounding.

6.3 HOW THE ASSOCIATIONAL CRITERION FAILS

We will say that a criterion for no-confounding is *sufficient* if it never errs when it classifies a case as no-confounding and *necessary* if it never errs when it classifies a case as confounding. There are several ways that the associational criterion of Definition 6.2.2 fails to match the causal criterion of Definition 6.2.1. Failures with respect to sufficiency and necessity will be addressed in turn.

6.3.1 Failing Sufficiency via Marginality

The criterion in Definition 6.2.2 is based on testing each element of T individually. A situation may well be present where two factors, Z_1 and Z_2, jointly confound X and Y (in the sense of Definition 6.2.2) and yet each factor separately satisfies (U_1) or (U_2). This may occur because statistical independence between X and individual members of T does not guarantee the independence of X and groups of variables taken from T. For example, let Z_1 and Z_2 be the outcomes of two independent fair coins, each affecting both X and Y. Assume that X occurs when Z_1 and Z_2 are equal and that Y occurs whenever Z_1 and Z_2 are unequal. Clearly, X and Y are highly confounded by the pair $T = (Z_1, Z_2)$; they are, in fact, perfectly correlated (negatively) without causally affecting